Assignment 1 **Evan Nichols**
EECS 731 8/30/16
Prof. Beckage

Three Potential Datasets

1. College Scorecard Data
   (a) 1996-2015
   (b) ITEM HERE
   (c) Loan performance, student debt, Labor Market Outcomes, Higher education
   (d) Students who take out loans are taking longer to repay their debt.
   (e) Compare the length of payment plans across students over the 1996 - 2015 period. Use a linear or quadratic regression model to try and predict every 4th year of data.
2. Stanford Wordbank
   (a) Contains data from 59,802 children and 66,635 CDI administrations, across 17 languages and 31 instruments.
   (b) Vocabulary development, Language and Cognition
   (c) Language Trajectories, Vocabulary Norms, Semantic networks, growth curves
   (d) Words that are a single syllable in one language will see faster growth/adoption curves than their multi-syllable counterparts (e.g. "dog" in English compared to "perro" in Spanish and "kopek" in Turkish).
   (e) This problem falls into a discrete, unsupervised category because we are attempting to find structure based on number of syllables. Create linear regression models for multiple words' adoption curves and compare across different languages (more specifically, comparing words that have different syllable counts across different languages).
3. Hoop Math - College Basketball Play-by-play Statistics
   (a) 2011 - 2016
   (b) Sports, Performance Analysis
   (c) Shooting, Assists, Offensive/Defensive Transition Splits, Putbacks
   (d) Team chemistry can be quantified as a weighted linear combination of all regular season statistics.
   (e) Create a model for "team chemistry" based on the relative average age of the starters, team assist rates and average PPG. Use this chemistry model on previous 3 years of regular season play data.